Prix des logements et autocorrélation spatiale : une approche semi-paramétrique

Ibrahim Ahamada, Emmanuel Flachaire, Marion Lubat

Prix des Logements et Autocorrélation Spatiale : une Approche Semi-Paramétrique

Ibrahim Ahamada
Emmanuel Flachaire
Marion Lubat

Université Paris 1 Panthéon-Sorbonne

Septembre 2007

Résumé

Dans cette étude, nous estimons l'influence de certaines caractéristiques sur les prix des logements avec la méthode des prix hédoniques. Nous utilisons tout d'abord une approche classique, basée sur un modèle de régression paramétrique avec autocorrélation spatiale. Cette approche présente deux inconvénients : la forme fonctionnelle du modèle et la matrice de poids sont fixées *à priori*. Nous présentons ensuite une approche semi-paramétrique qui permet de pallier ces limites.

1 Introduction

Les logements possèdent des caractéristiques propres (surface, nombre de pièces, ancienneté, jardin, etc.) et une série de facteurs de voisinage (accessibilité, statut socio-économique, infrastructures, services, environnement, etc.). Une méthode couramment employée pour évaluer l'influence de telles caractéristiques sur les prix des logements est la méthode des prix hédoniques. Celle-ci revient à considérer un modèle de régression où la variable dépendante est le prix de vente et les variables explicatives sont les caractéristiques propres et de voisinage du logement (Gravel et al. 1997, Cavailhès 2005, Géniaux et Napoléone 2005). Dans cet article, nous utilisons la méthode des prix hédoniques pour évaluer l'influence sur les prix de certaines caractéristiques propres au logement ainsi que de la proximité d'un espace vert.

Dans un premier temps, nous utilisons une approche paramétrique, basée sur un modèle de régression log-linéaire avec des aléas indépendants. Puis, nous prenons en considération la présence d'une éventuelle autocorrélation spatiale dans les aléas. En effet, même si des facteurs de voisinage sont pris en compte dans le modèle, comme la proximité d'un espace vert, la totalité des effets liés à la localisation est souvent impossible à capturer. L'hypothèse d'aléas indépendants est alors contestable et des méthodes adaptées doivent être utilisées. Nous utilisons les modèles d'autocorrélation spatiale pour tester puis tenir compte de ce phénomène (Anselin 1988, Jayet 1993, LeGallo 2002).

L'approche paramétrique présente deux limites principales. Tout d'abord, la nature de la relation entre une variable explicative et la variable dépendante doit être spécifiée *a priori* (sous forme linéaire, quadratique ...). Dans notre application, nous caractérisons l'influence des espaces verts sur les prix des logements en introduisant dans le modèle une variable muette, égale à 1 si un espace vert est situé à moins de 200 mètres du logement, et à 0 sinon. Ensuite, nous testons la présence d'un effet résiduel de la localisation géographique sur les prix des logements. Une telle présence étant décelée, nous utilisons les modèles d'autocorrélation spatiale, avec une matrice de poids telle que seuls les logements les plus proches, situés à une moins de 200 mètres, sont pris en compte dans la spécification de la dépendance spatiale. Même si une analyse approfondie des données pourrait guider ces choix, ils restent arbitraires et d'autres types de relation pourraient être employés.

Dans un deuxième temps, nous utilisons une approche semi-paramétrique, permettant de pallier les deux principales limites de l'approche paramétrique. Le modèle employé fournit une estimation de la nature de l'influence des espaces verts sur les prix des logements, tout en tenant compte de l'autocorrélation spatiale. La nature de la relation entre les espaces verts et les prix des logements revêt une forme différente de celle retenue dans l'approche paramétrique : les prix des logements sont effectivement influencés par la présence d'un espace vert si celui-ci n'est pas éloigné de plus de 200 mètres, mais le prix diminue de manière linéaire avec l'éloignement. Par ailleurs, la formulation de la dépendance spatiale met en évidence le fait que l'impact de la localisation géographique sur les prix est fortement non-linéaire.

2 Approche paramétrique

L'approche classique consiste à utiliser un modèle de régression linéaire, en tenant compte d'une éventuelle autocorrélation spatiale des aléas. Dans cette section, nous présentons tout d'abord la spécification habituellement employée en pratique, le modèle log-linéaire, puis nous étudions la présence éventuelle d'autocorrélation spatiale dans les aléas du modèle.

Les données disponibles portent sur les prix des transactions de biens immobiliers dans la ville de Brest en 1995, en milliers de francs, ainsi que les caractéristiques propres et extérieures aux différents biens. La taille du logement est spécifiée par son type : STUDIO, T1, T1BIS, T2, T3, T4, T5, T6, T7, T8, T9. L'introduction de m catégories distinctes dans le modèle avec des variables muettes, égales à 0 ou 1, nécessite l'ajout de $m-1$ variables dans le modèle. L'une d'entre elle n'est pas utilisée pour éviter un problème de colinéarité parfaite avec la constante et sert de catégorie de référence. Nous choisissons de prendre pour catégorie de référence le logement de type studio et n'introduisons donc pas la variable STUDIO dans le modèle. La présence ou non d'un parking est caractérisée par la variable muette PARKING. La distinction entre un appartement et une maison est également pris en compte avec la variable muette MAISON. Les coordonnées spatiales de chacun des ilots sont disponibles, ainsi que la localisation des espaces verts de la ville. Cela nous permet de calculer la distance euclidienne d'un logement à l'espace vert le plus proche. Afin de prendre en compte l'influence des espaces verts sur le prix des logements, nous introduisons dans le modèle une variable muette, PROXV, égale à 1 si un espace vert se situe à moins de 200 mètres du logement et à 0 sinon. Pour une description détaillée des données, voir Flachaire et al. (2006).

2.1 Modèle log-linéaire

Une manière simple de spécifier une fonction de prix hédonique est d'utiliser une régression linéaire en utilisant comme variable dépendante la transformation logarithmique du prix des logements, appelé modèle log-linéaire :

$$\log p = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon \tag{1}$$

où $X_1, \ldots, X_k$ sont les variables explicatives du modèle, $\beta_1, \ldots, \beta_k$ les paramètres inconnus et ε les aléas. Dans ce modèle, pour de faibles variations de p et X_j, la valeur du coefficient β_j mesure la variation relative du prix, consécutive à un changement d'une unité de la caractéristique X_j. Autrement dit, la valeur

$$WTP_j = (X_j^* - X_j)\beta_j$$

mesure la variation relative du prix d'un logement qu'un individu est prêt à payer, pour voir la valeur X_j de la caractéristique j passer à X_j^*.

Les résultats de l'estimation du modèle (1) sont présentés dans le tableau 1 (colonne Modèle log-linéaire), les écart-types étant en italique. Les aléas sont supposés indépendants et identiquement distribués (*i.i.d.*), le modèle est estimé par Moindres Carrés

TAB. 1 – Modèles paramétriques

variables	Modèle log-linéaire coef.	e. t.		Modèle spatial SEM coef.	e. t.	
Intercept	4.783	*(0.048)*	***	4.792	*(0.047)*	***
T1	0.057	*(0.060)*		0.065	*(0.057)*	
T1BIS	0.200	*(0.086)*	***	0.268	*(0.082)*	***
T2	0.375	*(0.054)*	***	0.368	*(0.051)*	***
T3	0.747	*(0.052)*	***	0.745	*(0.049)*	***
T4	0.941	*(0.053)*	***	0.968	*(0.050)*	***
T5	1.130	*(0.058)*	***	1.120	*(0.055)*	***
T6	1.311	*(0.074)*	***	1.317	*(0.069)*	***
T7	1.540	*(0.116)*	***	1.527	*(0.109)*	***
T8	1.583	*(0.148)*	***	1.576	*(0.138)*	***
T9	1.416	*(0.244)*	***	1.350	*(0.232)*	***
MAISON	0.263	*(0.031)*	***	0.224	*(0.030)*	***
PARKING	0.219	*(0.023)*	***	0.222	*(0.023)*	***
PROXV	0.105	*(0.025)*	***	0.078	*(0.027)*	***

Significativité : à 1% '***', à 5% '**', à 10% '*'.

Ordinaires (MCO). Ces résultats suggèrent par exemple que la valeur du prix moyen d'un T2 est 37.5% supérieure à la valeur du prix moyen d'un STUDIO, toute autre chose étant égale par ailleurs (coef=0.375). Par ailleurs, la présence d'un espace vert à proximité du logement conduit à une augmentation statistiquement significative du prix moyen d'un logement, de l'ordre de 10.5%.

Une spécification plus flexible peut être utilisée, le modèle Box-Cox, dont la formulation est comme suit :

$$g(p) = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon \quad \text{où} \quad g(p) = \begin{cases} (p^\lambda - 1)/\lambda & \text{pour } \lambda \neq 0 \\ \log p & \text{pour } \lambda = 0 \end{cases}$$

Le modèle linéaire correspond au cas où $\lambda = 1$ et le modèle log-linéaire correspond au cas où $\lambda = 0$. Aussi, le modèle Box-Cox permet de tester le modèle linéaire contre le modèle log-linéaire, alors que ces deux modèles ne sont pas emboités. Si aucun des deux modèles n'est retenu ($\lambda \neq 0, 1$), les résultats de l'estimation du modèle Box-Cox sont difficilement exploitables : (1) non seulement le coefficient β_j ne mesure pas une variation relative du prix car ce dernier n'est pas une transformation logarithmique du prix ; (2) mais aussi, le prix moyen d'un logement à caractéristiques données ne peut pas être calculé car la variable dépendante est une transformation non-linéaire du prix[1].

Les résultats de l'estimation du modèle Box-Cox étant difficiles à interpréter si les modèles linéaire et log-linéaire ne sont pas retenus ($\lambda \neq 0, 1$), la question qui se pose est de savoir ce que l'on fait dans un tel cas. Pour répondre à cette question, il est utile

[1] À partir du moment où la fonction $g(.)$ est une transformation non-linéaire du prix, on a $E[g(p)|.] \neq g(E[p|.])$, qui peut se réecrire : $\hat{E}[p|X_1, \ldots, X_k] = g^{-1}(E[g(p)|X_1, \ldots, X_k]) + \text{biais}$. Le problème est que le biais est en général inconnu, il ne s'atténue pas lorsque la taille de l'échantillon augmente.

de regarder les propriétés du paramètre λ. En fait, le paramètre de transformation λ joue deux rôles distincts : il affecte la forme fonctionnelle du modèle mais également les propriétés du terme d'erreur. Par exemple, l'estimation d'un modèle Box-Cox à partir de données générées par un modèle linéaire hétéroscédastique conduirait souvent à une estimation de λ inférieure à 1, afin de diminuer le montant de l'hétéroscédasticité (Davidson et MacKinnon 1993, section 14.7). On serait ainsi amené à conclure incorrectement que la forme linéaire n'est pas appropriée. Notons que le problème de l'hétéroscédasticité est fréquent dans les données individuelles.

L'analyse précédente nous conduit à recommander la méthodologie suivante :

1. L'estimation du modèle Box-Cox permet de tester un modèle linéaire ($\lambda = 1$) vs. un modèle semi-log ($\lambda = 0$), qui en sont des cas particulier. Si l'une des deux hypothèses n'est pas rejetée, cela permet de sélectionner l'un des deux modèles.
2. Une estimation de λ significativement différente de 0 et 1 peut être due à la présence d'hétéroscédasticité dans le modèle. On sélectionne le modèle linéaire si λ est plus proche de 1 que de 0 et le modèle semi-log sinon, puis on traite le problème de l'hétéroscédasticité pour le modèle sélectionné.

Avec nos données, l'estimation du modèle Box-Cox conduit à sélectionner le modèle log-linéaire présenté précédemment.

2.2 Autocorrélation spatiale

Le prix d'un logement n'est pas seulement le fruit d'une combinaison d'attributs qui lui sont propres. Les logements ont une localisation géographique particulière qui peut avoir un effet important dans la détermination des prix. Une manière de capturer ce type d'effet consiste à introduire des régresseurs dans le modèle. Si elles sont disponibles, des mesures telles que l'accessibilité au centre ville (distances ou proximité au centre ville, à l'autoroute, au métro etc) ou la qualité du voisinage (taux d'échec scolaire, de chômeurs, proximité à un espace vert, présence d'une gare etc.) peuvent être utilisées. Néanmoins, même si de telles variables sont prise en compte dans le modèle, il est difficile de capturer complètement l'effet de la localisation géographique sur les prix. Dans notre application, nous avons pris en compte la présence d'un espace vert à proximité du logement. Il est fort probable que d'autres variables liées à la localisation géographique aient un effet sur le prix des logements.

Si l'effet de la localisation sur les prix n'est pas complètement pris en compte par l'ajout de régresseurs dans le modèle, un effet résiduel devrait persister dans le terme d'erreur du modèle, se traduisant par une dépendance ou autocorrélation spatiale. L'hypothèse d'indépendance du terme d'erreur n'est alors plus vérifiée et la méthode d'estimation par les Moindres Carrés n'est pas appropriée. L'approche traditionnelle utilisée pour traiter ce phénomène est développée dans le cadre de l'autocorrélation spatiale. La dépendance spatiale est prise en compte par le biais d'une matrice de poids W, qui spécifie les positions relatives des observations les unes par rapport aux autres. Cette matrice est exogène, elle est définie *a priori* par le modélisateur. Les éléments diagonaux

w_{ii} sont égaux à 0 tandis que les éléments non-diagonaux w_{ij} indiquent comment l'unité i est spatialement connectée à l'unité j. Ces éléments sont non-négatifs et finis. La matrice de poids est standardisée de telle manière que la somme des éléments d'une même ligne soit égale à 1. Divers types de structure spatiale peuvent être utilisés (méthodes des contiguités, voisins les plus proches et fonction de distance finie). Comme il est rare qu'un type s'impose a priori comme le meilleur, il faut souvent tester plusieurs choix avant d'en sélectionner un. Pour une discussion détaillée sur l'autocorrélation spatiale, voir entre autres Anselin (1988), Can (1992), Jayet (1993, 2001) et LeGallo (2002, 2004).

Pour tester la présence d'autocorrélation spatiale, le test le plus courant est le test I de Moran. La logique du test I de Moran est une simple extension du test d'autocorrélation des résidus pour les données chronologique, au cas de deux dimensions (données géographiques). Il s'écrit comme suit :

$$I = \frac{\hat{\varepsilon}^\top W \hat{\varepsilon}}{\hat{\varepsilon}^\top \hat{\varepsilon}} \quad \sim \quad N(0,1)$$

où $\hat{\varepsilon}$ est le vecteur des résidus du modèle initial estimé par MCO. Ce test est sensible au choix *a priori* de la matrice des poids W.

Si le test I de Moran rejette l'hypothèse nulle d'absence d'autocorrélation spatiale, la dépendance spatiale doit être prise en compte dans le modèle de régression initial. Deux modèles de référence peuvent être employés : les modèles spatiaux SEM (*spatial error model*) et LAG (*spatial lag model*) :

- Le Modèle SEM suppose que le terme d'erreur est spatialement dépendant. L'autocorrélation spatiale est alors modélisée avec un terme d'erreur qui suit un processus spatial autorégressif :

$$\log p = X\beta + u \qquad u = (Wu)\lambda + \varepsilon \tag{2}$$

  où ε est un bruit blanc et X est une matrice composée de variables explicatives. La détection d'autocorrélation spatiale dans le terme d'erreur indique souvent un problème de spécification du modèle, telle que l'omission de variables explicatives. L'effet spatial, qui n'est pas complètement capturé par les régresseurs se répercute alors dans le terme d'erreur.

- Le Modèle LAG suppose implicitement que la moyenne pondérée des prix des logements voisins affecte le prix d'un logement. L'autocorrélation spatiale est alors modélisée en introduisant les prix des logements voisins dans les régresseurs :

$$\log p = X\beta + (W \log p)\,\rho + \varepsilon \tag{3}$$

  où ρ est un paramètre spatial autorégressif, indiquant l'ampleur de l'intéraction existante entre les observations, et ε est un bruit blanc. Dans ce modèle, l'observation $\log p_i$ est en partie expliquée par les valeurs prises par $\log p$ dans les régions voisines : $[W \log p]_i = \sum_{j \neq i} w_{ij} \log p_i$. La matrice W étant standardisée, cette valeur s'interprète comme la moyenne des valeurs de $\log p$ sur les observations voisines à i.

Le choix d'un modèle plutôt qu'un autre se fait à l'aide de tests de spécification. Dans un premier temps, on peut utiliser des statistiques de tests LM standards :

- LMerr permet de tester l'hypothèse nulle $H_0 : \lambda = 0$ à partir du modèle (2).
- LMlag permet de tester l'hypothèse nulle $H_0 : \rho = 0$ à partir du modèle (3),

Dans le cas où les deux tests conduisent au rejet de l'hypothèse nulle, cela conduit à suspecter un problème d'autocorrélation spatiale, mais cela ne permet pas de sélectionner l'un des deux modèles, LAG ou SEM. On est alors amené, dans un deuxième temps, à tester la présence d'autocorrélation spatiale à partir d'un modèle plus général. Les deux modèles SEM et LAG peuvent en effet être combiné, ils sont des cas particulier du modèle suivant :

$$\log p = X\beta + (W \log p)\,\rho + u \qquad u = (Wu)\lambda + \varepsilon \tag{4}$$

On teste la présence d'autocorrélation spatiale d'une certaine forme, robuste à la présence d'une autre forme :

- RLMerr permet de tester l'hypothèse nulle $H_0 : \lambda = 0$ à partir du modèle (4).
- RLMlag permet de tester l'hypothèse nulle $H_0 : \rho = 0$ à partir du modèle (4).

Le non-rejet de l'hypothèse nulle par l'un des deux tests permet de sélectionner un modèle.

Avec nos données, nous choisissons de définir la structure spatiale avec la méthode des voisins les plus proches et en identifiant les logements situés dans un ilôt à moins de d mètres de l'ilôt du logement de référence. La valeur $d = 200$ est utilisée, c'est-à-dire qu'un logement est corrélé spatialement avec les logements qui l'entourent, dans un périmètre de 200 mètres[2]. Les résultats du test I de Moran à partir du modèle semi-log sont présentés dans le tableau ci-dessous :

I de Moran	Espérance	Variance	P-value
0.175940	-0.001744	0.000349	<2.2e-16

L'hypothèse nulle d'absence d'autocorrélation spatiale est rejetée sans ambiguité. Il apparaît donc que le modèle considéré jusqu'à présent est spatialement corrélé. Il faut maintenant estimer le modèle en tenant compte de ce problème.

Afin de pouvoir éventuellement choisir entre un modèle SEM et LAG, nous calculons les statistiques de tests LM, dont les valeurs sont présentées dans le tableau suivant :

	LMerr	LMlag	RLMerr	RLMlag
statistique	117.09	5.89	113.13	1.93
P-value	<2.2e-16	0.015	<2.2e-16	0.164
significativité	***	**	***	

[2]D'autres choix auraient pu être utilisés. Notamment, une étude approfondie des données et des dispositions géographiques des transactions sont susceptibles de guider la spécification de la matrice de poids, cette dernière devant être construite pour refléter la structure de connexion entre les localisations de transactions. Concernant notre choix, même si une telle étude n'a pas été menée, nous verrons qu'il est approprié à partir de l'analyse semi-paramétrique présentée dans la section suivante.

Sur la base des tests LMerr et LMlag, l'absence d'autocorrélation spatiale est rejetée, mais il n'est pas évident de sélectionner un modèle plutôt qu'un autre. L'emploi des tests robustes RLMerr et RLMlag permet d'en dire plus. La statistique de test RLMerr conduit à rejetter l'hypothèse nulle : en présence ou non de retard de la variable endogène, la présence d'autocorrélation spatiale dans le terme d'erreur est significative. D'un autre côté, la statistique RLMlag conduit à ne pas rejetter l'hypothèse nulle : en présence ou non d'un terme d'erreur spatialement dépendant, la présence d'un retard de la variable endogène n'est pas significative. Ces résultats nous conduisent à sélectionner le modèle SEM dans la suite.

La sélection du modèle SEM dans notre contexte n'est pas surprenant. Le modèle SEM est en effet beaucoup plus vraisemblable que le modèle LAG, car le fait que des variables explicatives ayant une structure spatiale soient non disponibles, et donc incorporées dans le terme d'erreur, est quasiment une évidence. Par contre, le modèle LAG suppose que chaque transaction est influencée directement par les transactions voisines, ce qui ne peut guère se comprendre que par un effet d'information qu'il est difficile à justifier, et qui n'est valable que pour les transactions voisines antérieures à la vente du bien considéré.

Les résultats de l'estimation du modèle SEM, avec la fonction `errorsarlm` de la librairie `spdep` du logiciel R, sont présentés dans le tableau 1. Les coefficients qui évaluent l'influence des espaces verts sur les prix des logements sont sensiblement différents dans le modèle log-linéaire avec aléas *i.i.d.* (coef=0.105) et dans le modèle spatial (coef=0.078). Si on suspecte que des variables de localisation sont omises dans le modèle initial, un tel résultat n'est pas surprenant. En effet, l'omission de certaines variables dans le modèle a notamment pour conséquences de :

- générer un biais sur les coefficients des variables corrélées avec les variables omises,
- ne pas générer de biais pour les variables indépendantes avec les variables omises.

Les variables de distance et de voisinage sont susceptibles d'être fortement corrélées aux coordonnées géographiques. Ce qui suggère que le coefficient de la variable PROXV du modèle log-linéaire, estimé par MCO, est biaisé : il capture des effets de localisation ou de voisinage autres que ceux de l'influence des espaces verts sur les prix des logements. Dans notre cas, l'effet des espaces verts sur les prix des logements est sur-estimé dans le modèle initial. La prise en compte de l'autocorrélation spatiale permet de mettre en évidence et de corriger une telle sur-évaluation.

3 Approche semi-paramétrique

L'approche paramétrique présentée jusqu'ici présente deux limites principales : la variable PROXV et la matrice des poids W sont fixées de manière arbitraire. Autrement dit, l'influence des espaces verts sur les prix des logement et la nature de la dépendance spatiale sont spécifiées *a priori*, d'autres choix pouvant être possibles. Dans cette section, nous présentons un approche semi-paramétrique qui permet de pallier ces deux limites.

Afin de relâcher la spécification *a priori* de l'influence des espaces verts sur les prix des logements faite par la variable PROXVERT, nous utilisons directement la distance euclidienne à l'espace vert le plus proche (DISTV) comme régresseur, sans spécifier la relation qui la lie à la variable dépendante. Le modèle de régression s'écrit sous la forme d'un modèle partiellement linéaire :

$$\log p = X'\beta + s_1(\text{DISTV}) + \varepsilon$$

où la matrice X' contient les régresseurs du modèle log-linéaire sauf la variable PROXV. La fonction $s_1(.)$ est inconnue, elle est estimée par des méthodes semi-paramétriques. Le modèle log-linéaire de la section précédente est un cas particulier de ce modèle. En effet, la variable PROXV équivaut à la forme de la fonction $s_1(.)$ suivante :

$$s_1(\text{DISTV}) = \begin{cases} 1 & \text{si DISTV} < 200 \\ 0 & \text{sinon} \end{cases}$$

Remarquons que les régresseurs dans la matrice X' sont des variables muettes - égales à 0 ou 1 - pour lesquelles une spécification quelconque n'aurait pas de sens.

Avec le même raisonnement, il est également possible de relâcher la spécification *a priori* de la matrice de poids W. Cette matrice tiend compte du fait que la localisation géographique - mesurée par les deux coordonnées C_1 et C_2 - peut avoir une influence sur les prix des logements. Une telle relation peut être spécifiée en considérant le modèle de régression semi-paramétrique suivant :

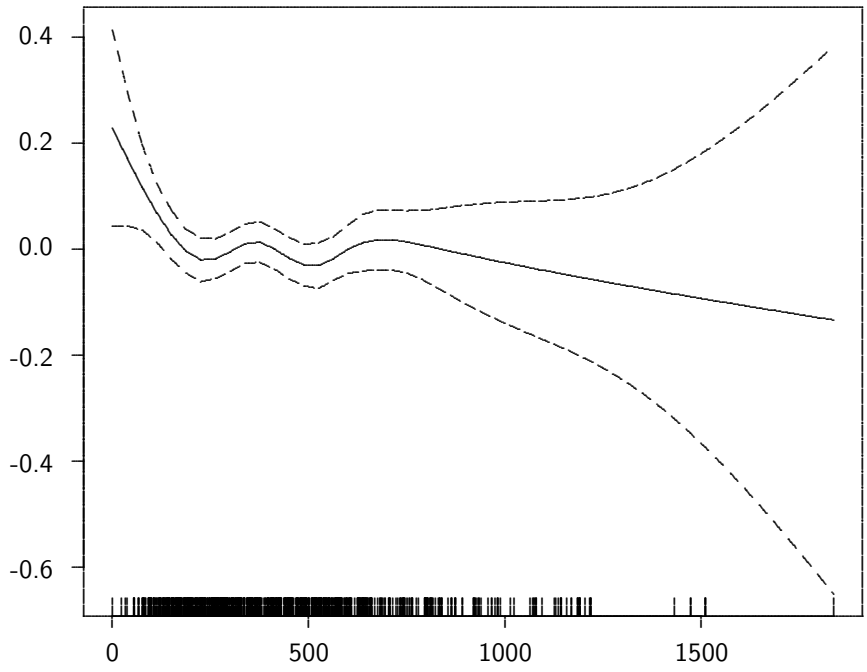
$$\log p = X'\beta + s_1(\text{DISTV}) + s_2(C_1, C_2) + \varepsilon \tag{5}$$

Les coordonnées géorgraphiques apparaissent maintenant dans le modèle en tant que régresseurs et leur influence sur la variable dépendante est exprimée par la fonction inconnue $s_2(.)$.

L'estimation d'un modèle semi-paramétrique de ce type n'est pas sans difficultés. Une première difficulté porte sur le choix du paramètre de lissage. Dans notre application, nous estimons ce modèle avec la fonction `gam` de la librairie `mgcv` du logiciel R, qui utilise une méthode de sélection automatique du paramètre de lissage. Une deuxième difficulté, connue sous le nom de "fléau de la dimension", suggère qu'il faut beaucoup de données pour estimer la fonction $s_2(.)$ avec suffisament de précision. Dans notre application, nous disposons d'un grand nombre d'observations ($n = 1157$). Pour plus d'informations sur l'estimation des modèles semi-paramétriques, les ouvrages de Pagan et Ullah (1999), Yatchew (2003) et Ahamada et Flachaire (2008) peuvent être consultés.

La figure 1 présente l'estimation de la fonction $s_1(.)$ à partir de l'estimation du modèle (5). Ce graphique met en évidence la relation entre le prix des logements et la distance à l'espace vert le plus proche. La trait plein correspond à l'estimation de la fonction $s_1(.)$, avec pour abscisse la distance du logement à l'espace vert le plus proche. Les traits en pointillés déterminent un intervalle de confiance à 95%. Ce graphique montre que la proximité immédiate d'un espace vert a un effet positif. Lorsqu'on s'éloigne

FIG. 1 – Estimation de la fonction s_1(DISTV)



du parc, l'effet chute très vite, de façon linéaire, jusqu'à ce qu'il devienne nul à une distance approximative de 200 mètres. Au-delà, la relation n'est pas significativement différente de zéro.

L'effet des espaces verts sur les prix des logements suggéré par le graphique 1 - relation linéaire décroissante jusqu'à une distance d'environ 200 mètres et effet nul ensuite - peut être modélisé de manière paramétrique au moyen d'une relation linéaire fragmentée (voir Gujarati 2004, section 9.8). Cette modélisation consiste à introduire une variable muette dans le modèle initial :

$$\log p = X'\beta + \gamma_1 \text{DISTV} + \gamma_2(\text{DISTV} - \text{DISTV}^*)M + s(\text{C}_1, \text{C}_2) + \varepsilon \qquad (6)$$

où DISTV* est le seuil, appelé aussi noeud, égal à 200 dans notre cas. La variable M est une variable muette égale à 1 si la distance est supérieure au seuil (DISTV $>$ DISTV*) et à 0 sinon. Autrement dit, si la distance du logement à l'espace vert le plus proche est inférieure ou égale au seuil, DISTV $\leq$ DISTV*, le modèle devient :

$$\log p = X'\beta + \gamma_1 \text{DISTV} + s(\text{C}_1, \text{C}_2) + \varepsilon$$

Le paramètre γ_1 mesure la variation relative du prix d'un logement (en %), tout autre chose étant égale par ailleurs, pour un éloignement à l'espace vert le plus proche de 1 mètre par rapport à sa position initiale. Si la distance du logement à l'espace vert le plus proche est supérieure au seuil, DISTV $>$ DISTV*, le modèle devient :

$$\log p = X'\beta + \gamma_2 \text{DISTV}^* + (\gamma_1 + \gamma_2)\text{DISTV} + s(\text{C}_1, \text{C}_2) + \varepsilon$$

TAB. 2 – Modèles semi-paramétrique vs. paramétriques

	Modèle semi-param.			Modèle spatial SEM			Modèle log-linéaire		
variables	coef.	e. t.		coef.	e. t.		coef.	e. t.	
Intercept	4.992	*(0.084)*	***	5.015	*(0.088)*	***	5.136	*(0.086)*	***
T1	0.075	*(0.055)*		0.069	*(0.057)*		0.057	*(0.060)*	
T1BIS	0.245	*(0.080)*	***	0.267	*(0.082)*	***	0.196	*(0.086)*	**
T2	0.381	*(0.050)*	***	0.369	*(0.051)*	***	0.371	*(0.054)*	***
T3	0.763	*(0.049)*	***	0.747	*(0.049)*	***	0.746	*(0.052)*	***
T4	0.998	*(0.050)*	***	0.968	*(0.050)*	***	0.939	*(0.053)*	***
T5	1.154	*(0.055)*	***	1.120	*(0.055)*	***	1.126	*(0.058)*	***
T6	1.350	*(0.069)*	***	1.320	*(0.069)*	***	1.311	*(0.074)*	***
T7	1.599	*(0.108)*	***	1.536	*(0.109)*	***	1.544	*(0.116)*	***
T8	1.569	*(0.137)*	***	1.588	*(0.138)*	***	1.593	*(0.147)*	***
T9	1.468	*(0.227)*	***	1.352	*(0.232)*	***	1.415	*(0.244)*	***
MAISON	0.230	*(0.031)*	***	0.221	*(0.030)*	***	0.263	*(0.031)*	***
PARKING	0.224	*(0.023)*	***	0.217	*(0.023)*	***	0.215	*(0.023)*	***
DISTV	-0.0011	*(0.0003)*	***	-0.0011	*(0.0004)*	***	-0.0017	*(0.0004)*	***
(DISTV-DISTV*)M	0.0011	*(0.0003)*	***	0.0011	*(0.0004)*	***	0.0017	*(0.0004)*	***

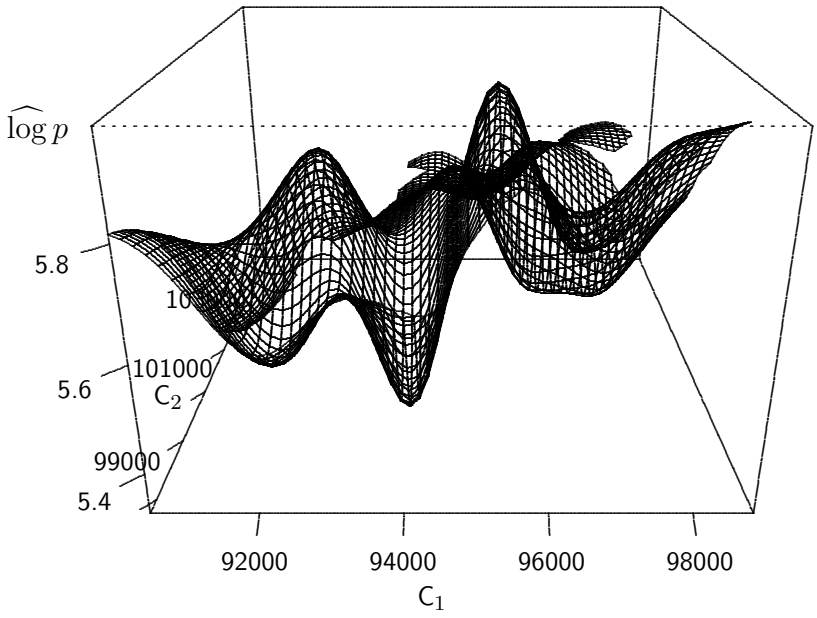
Significativité : à 1% '***', à 5% '**', à 10% '*'.

Le paramètre $\gamma_1 + \gamma_2$ mesure la variation relative du prix d'un logement (en %), tout autre chose étant égale par ailleurs, pour un éloignement à l'espace vert le plus proche de 1 mètre par rapport à sa position initiale. Il est clair que pour les deux cas de figures, les coefficients associés à la variable DISTV sont différents : l'effet de cette variable sur le prix des logements n'est pas le même selon le cas où l'on se trouve. Pour refléter la relation suggérée par la figure 1, les coefficients γ_1 et γ_2 du modèle (6) devraient être comme suit :

- si la distance est inférieure à 200 mètres, l'influence est décroissante : $\gamma_1 < 0$,
- si la distance est supérieure à 200 mètres, l'influence est inexistante : $\gamma_1 + \gamma_2 = 0$.

Le tableau 2 présente les résultats de l'estimation du modèle semi-paramétrique (6), ainsi que des modèles log-linéaire et spatial SEM présentés dans la section précédente. Pour que la comparaison soit possible avec le modèle semi-paramétrique, la variable PROXV est remplacée par les variables DISTV et (DISTV-DISTV*)M dans les modèles paramétriques (log-linéaire et spatial SEM). Les résultats sont cohérents avec l'analyse précédente : le coefficient de la variable DISTV est significativement négatif ($\hat{\gamma}_1 = -0.0011$) et la somme des coefficients des variables DISTV et (DISTV-DISTV*)M n'est pas significativement différente de zéro ($\hat{\gamma}_1 + \hat{\gamma}_2 \approx 0$). Cela indique que, pour les logements situés à moins de 200 mètre d'un espace vert, plus on s'éloigne du parc plus le prix diminue. La différence de prix moyen entre un logement à proximité immédiate d'un espace vert et un autre éloigné de 100 mètres, tout autre chose étant égale par ailleurs, est de l'ordre de 11%. Par ailleurs, l'hypothèse $\gamma_1 + \gamma_2 = 0$ n'étant pas rejetée, les espaces verts n'ont pas d'influence sur les prix des logements qui sont éloignés de plus de 200 mètres. Si maintenant on compare les modèles entre eux, on constate que le modèle semi-paramétrique et le modèle spatial SEM ont des résultats assez proches. Par

FIG. 2 – Estimation de la fonction $s(C_1, C_2)$



contre, le modèle log-linéaire fournit des résultats sensiblement différents. Par exemple, le coefficient de la variable DISTV est égal à -0.0017, ce qui suggère que la différence de prix moyen entre un logement à proximité immédiate d'un espace vert et un autre éloigné de 100 mètres, tout autre chose étant égale par ailleurs, est de l'ordre de 17%. Comme nous l'avons constaté dans la section précédente, ce résultat indique une sur-évaluation de l'influence des espaces verts sur les prix des logements lorsque la dépendance spatiale n'est pas prise en compte.

L'estimation du modèle semi-paramétrique (6) fournit également une estimation de la fonction $s(.)$, permettant d'évaluer l'impact de la localisation géographique sur les prix des logements. La figure 2 présente l'estimation de cette fonction, dans un représentation à trois dimensions (C_1, C_2 et $\widehat{\log p}$). Une très forte non-linéarité de la fonction $s(.)$ est mise en évidence. Cette non-linéarité reflète les caractéristiques géographiques de la ville de Brest. Deux pics sont situés aux alentours de C_1 égal à 9300 et 9500 avec un creux entre les deux. Le creux correspond à la localisation de la rivière qui traverse une partie de la ville, la Penfeld. Le pic les plus élevé ($C_1 \approx 9500$) correspond au quartier du centre ville (autour de l'hôtel de la mairie). Le pic un peu moins élevé se situe de l'autre côté de la Penfeld. Autrement dit, nous détectons deux localisations où les logements sont les plus chers, toute autre chose égales par ailleurs, situées de part et d'autre de la Penfeld.

4 Conclusion

Dans cet article, nous avons utilisé une approche paramétrique et semi-paramétrique pour évaluer l'influence de certaines caractéristiques sur les prix des logements. Plusieurs résultats peuvent être soulignés. Tout d'abord, l'approche semi-paramétrique nous a permis de justifier une formulation de l'influence des espaces verts sur les prix des logements : linéaire et décroissante jusqu'à une distance de 200 mètres environ. Ensuite, l'approche semi-paramétrique a mis en évidence un effet fortement non-linéaire de la localisation géographique sur les prix, détectant deux zones où les prix des logements sont les plus élevés. Finalement, la comparaison des résultats de l'estimation des différents modèles a montré que la non-prise en compte de l'autocorrélation spatiale dans les aléas conduisait à sur-évaluer l'influence des espaces verts sur les prix des logements.

Références

Ahamada, I. et E. Flachaire (2008). *Économétrie Non-Paramétrique*. Economica, à paraître.

Anselin, L. (1988). *Spatial Econometrics : Methods and Models*. Springer.

Can, A. (1992). "Specification and estimation of hedonic housing price models". *Regional Science and Urban Economics*, **22**, 453–474.

Cavailhès, J. (2005). "Le prix des attributs des logements". *Économie et Statistique*, **381**, 91–123.

Davidson, R. et J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York : Oxford University Press.

Flachaire, E., H. Jayet, L. Ragot, et J. P. Tropéano (2006). *Économie Urbaine et Espaces Verts*. Rapport de recherches contractuelles du programme "Science Economique et Environnement" APR S3E 2002 du Ministère de l'Environnement.

Géniaux, G. et C. Napoléone (2005). "Rente fonciere et anticipations dans le periurbain". *Économie et Prévision*, **168**, 77–95.

Gravel, N., M. Martinez, et A. Trannoy (1997). "Une approche hédonique du marché des logements". *Études Foncières*, **74**.

Gujarati, D. N. (2004). *Économétrie*. Bruxelles : DeBoeck.

Halvorsen et H. O. Pollaskowski (1981). "Choice of functional form for hedonic price equations". *Journal of Urban Economics*, **10**, 37–49.

Jayet, H. (1993). *Analyse Spatiale Quantitative : Une Introduction*. Paris : Economica.

Jayet, H. (2001). "Econométrie et données spatiales". *Cahiers d'Économie et Sociologie Rurales*, **58-59**, 105–129.

LeGallo, J. (2002). "Econométrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaires". *Economie et Prévision*, **155**, 139–158.

LeGallo, J. (2004). "Hétérogénéité spatiales : principes et méthodes". *Economie et Prévision*, **162**, 151–172.

Pagan, A. et A. Ullah (1999). *Nonparametric Econometrics*. Cambridge : Cambridge University Press.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge : Cambridge University Press.